IMPLICIT COMMUNICATION OF WORD MEANING THROUGH CO-OCCURRENCE

Jeroen van Paridon[*1] and Gary Lupyan[*1]

[*]Corresponding Author: vanparidon@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, USA

Communication requires people to align, at least in part, on what words mean. To accomplish this, language learners can observe what a word refers to. However, many words have referents that are abstract or otherwise hard to observe. In addition, many concrete words also have more abstract denotations. In the absence of direct referents, language learners can align word meanings to those of their language community by observing how words are used in context. This mechanism is underappreciated as a driver of semantic alignment across large language communities in which individual speakers are unlikely to ever interact directly. In three experiments, we demonstrate alignment between blind and sighted speakers for the semantic associations of color terms – whose direct referents can only be observed by the sighted participants – and demonstrate how a word embedding model can achieve this alignment by learning from word co-occurrence patterns.

Experiments in language evolution often focus on the transmission of structure rather than the transmission of semantics. For example, Kirby, Cornish, and Smith (2008) write that "each utterance has a dual purpose, carrying semantic content but also conveying information about its own construction. Upon hearing a sentence, a language learner uses the structure of that sentence to make new inferences about the language that produced it". This is a foundational claim in the field of language evolution, and it has been repeatedly demonstrated empirically through, e.g., iterated learning experiments. In addition to conveying pragmatic meaning and information about its own construction, however, language also carries implicit information about its own lexical semantics in the form of word co-occurrences. For example, we might learn that "odd", "strange", and "weird" are related because because they are used in similar contexts. This is an underappreciated mechanism for language evolution, as it serves to align lexical semantics across speakers in a language community, which is vital for developing and maintaining a mutually intelligible lexicon.

Of course learning what words mean involves more than than tracking co-occurrences. Often, there is a direct referent present that the learner can observe. In some cases however, the language itself is the only source of information about lexical semantics that a language learner has access to. Blind people, for example,

can only learn about the meanings of color words through language. We would expect, therefore, that they acquire those aspects of color word semantics that are implicitly conveyed in spoken and written language.

Recently, Saysani, Corballis, and Corballis (2021) showed that blind people's judgments resemble those of sighted people when asked to place color words along various dimensions, for example indicating where "red" and "green" fall on cold-hot, unripe-ripe and fast-slow continua. Given that blind people cannot directly observe that hot objects sometimes glow red or that unripe fruits and vegetables tend to be green, it perhaps seems obvious that any color associations they do have, they must learn from language (cf. Kim, Aheimer, Manrara, & Bedny, 2021). However, *how* color semantics are represented in spoken and written language – and to what extent language, rather than perception, can align semantic representations of colors between individuals – is not obvious. Are color semantics conveyed explicitly, e.g. through generic statements such as "green fruits are unripe"? Are they conveyed through simple co-occurrences, when a color word occurs adjacent to another word, e.g. "red hot coals"? Or are color semantics encoded in more complex semantic structures – a web of associations from which we can derive semantics of color terms?

Experiment 1: Reanalysis of Saysani et al. (2021) data

Method

Participants

Saysani et al. recruited 32 native speakers of New Zealand English, 20 of whom had normal, trichromatic vision and 12 of whom were congenitally blind with no residual vision. We recruited 130 additional sighted participants from the student participant pool at a large public university, speakers of American English.

Design and procedure

Participants were asked to rate each of nine color terms (red, orange, yellow, green, blue, brown, purple, black, and white) on 17 semantic dimensions, each defined by two antonyms placed at the poles of a seven-point Likert scale (happy–sad, calm–angry, submissive–aggressive, relaxed–tense, exciting–dull, selfless–jealous, active–passive, like–dislike, alive–dead, fast–slow, new–old, unripe–ripe, soft–hard, light–heavy, fresh–stale, clean–dirty, and cold–hot).

Results

The main finding reported by Saysani et al. was that multidimensional scaling solutions were more variable between blind participants than between sighted participants. When we compared intraclass correlations (ICC) for the blind (.35, 95% CI [.29, .42]) and the sighted (.49, 95% CI [.43, .55]) groups, blind participants

were indeed more variable than the sighted participants. At the same time, the responses of sighted and blind participants were remarkably similar (see Figure 1).

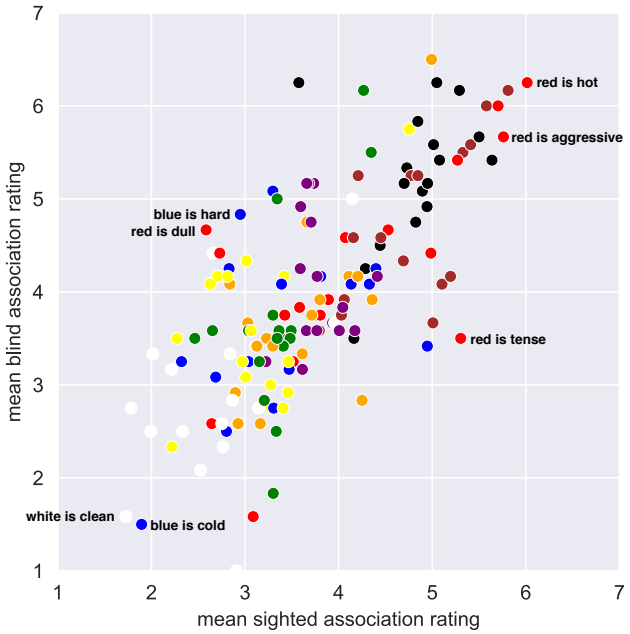

Figure 1. Blind and sighted participants' color-adjective association ratings from Experiment 1. Points on the diagonal from bottom-left to top-right represent perfect agreement between blind and sighted participants.

To understand how this aspect of color semantics may be represented in language, we relied on a fastText word embedding model (Bojanowski, Grave, Joulin, & Mikolov, 2017) trained on the fiction subcorpus of the Corpus of Contemporary American English (COCA-fiction). We projected the vector-representation of each color word onto a semantic dimension formed by the antonym pairs, e.g., *hot* and *cold*, and computing the cosine similarity between the color word vector and the axis vector. The projection for e.g. the color blue on the dimension cold-hot is then given by $\cos\left(\overrightarrow{\text{hot}} - \overrightarrow{\text{cold}}, \overrightarrow{\text{blue}}\right)$ (see Grand, Blank, Pereira, & Fedorenko, 2018, for a discussion of this projection method). This provides us with a relative measure of word similarity, taken along the semantic dimension's axis, that we can use to predict human ratings of color associations.

Using a Bayesian linear mixed-effects model with weakly regularizing pri-

ors (Capretto et al., 2020), we regressed word embedding projections onto participants' color-adjective association ratings while adjusting for frequency and concreteness of the words forming each dimension. Color-adjective ratings (e.g., placing yellow closer to ripe than unripe) were predicted by word embedding projections, with a standardized effect size of .40 (95% CI [.37, .43]) for sighted participants and .33 (95% CI [.24, .41]) for blind participants.

Discussion

Language is produced by people, most of whom have direct experiences of color. What is remarkable however, is that color information then becomes embedded in the statistics of language, enabling – in principle – someone who has no direct experience of color whatsoever to build up meaningful color semantics that can produce judgments quite similar to that of sighted people.

Experiment 2: Where in language are color associations coming from?

So where do the embeddings "learn" their color semantics? One way of finding out is to remove the critical signal from the training corpus so that the resulting word embeddings no longer predict human judgments. In this experiment, we examined four potential sources of color-adjective associations:

(a) *First-order* co-occurrences: The occurrence of a color word and a semantic dimension word in the same sentence (e.g. "the fire was *red hot*"; color associations in these sentences can be explicit, but often are not).

(b) *Second-order* co-occurrences: The occurrence of color words and semantic dimension words in similar contexts (i.e. color words and semantic dimension words may not co-occur, but share words that they co-occur with, e.g. "Southern cooking uses *green* tomatoes" and "Southern cooking uses *unripe* tomatoes"). These sentences encompass nearly the entire corpus because some words (e.g. many function words) co-occur with every other word, which made removing all of them from the training corpus infeasible. More importantly, it rules out a strong form of the second-order co-occurrence hypothesis (i.e. *all* second-order co-occurrence relationships are informative), but it does not preclude a weaker form, where *some* second-order co-occurrences (e.g. the psychologically salient words from hypothesis (d)) are central to learning color-adjective associations.

(c) Co-occurrences between color words and words in the same semantic neighborhood as semantic dimension words: For example in "The forest was *white* with *snow*", *snow* is in the same semantic neighborhood as *cold*, which might lead to an association between *white* and *cold*). We identified semantic neighborhood words using cosine similarity between word embeddings and removed sentences containing any of the ten nearest neighbors of each color and dimension word from the corpus.

(d) Mediation by psychologically salient words: It is possible that color-adjective associations are mediated by specific words. For example, when placing *yellow* on the *unripe*-to-*ripe* dimension, people may think of a yellow and ripe banana. We do not know *a priori* which words mediate color-adjective associations, but we presented participants with color-adjective pairs (e.g., yellow-ripe, white-cold) and asked them to provide a word they associate with the pair. We then take the most common word for each pair and remove sentences containing those words from the training corpus.

Note that these sources of semantic information need not be mutually exclusive; words captured by (c) and (d) may overlap, and all of these words may be a subset of the words described by (b).

Method

Participants

We recruited 100 sighted participants from the student participant pool at a large public university who did not participate in previous color-adjective rating studies. Participants were presented with the color-adjective pairs and asked to generate a word that they associate with both. These associates were taken to be psychologically salient mediator words from hypothesis (d).

Design and procedure

To test each potential source of color-adjective associations, we removed it from the training corpus and then tested the predictive efficacy of embedding projections trained on the filtered corpus by using them to model the association ratings from Experiment 1.

Results

Removing first-order co-occurrences did not meaningfully reduce the effect size of the word embedding predictions. Removing nearest neighbors and especially removing participant-generated labels for color-adjective associations had a measurable impact however (see Figure 2 for estimated effect sizes).

Discussion

It is tempting to think that knowledge that blue is cold may come from sentences such as "His lips were blue with cold". However, removing such first-order co-occurrences had no measurable effect on the model's ability to pull out human-like associations. In contrast, removing sentences containing psychologically salient mediators (e.g., "ice" for cold-blue) reduced the signal substantially. This is especially surprising because the number of labels generated by at least two participants (the threshold for inclusion in our corpus filtering procedure) was only 242; on average less than one label per color-adjective pair.

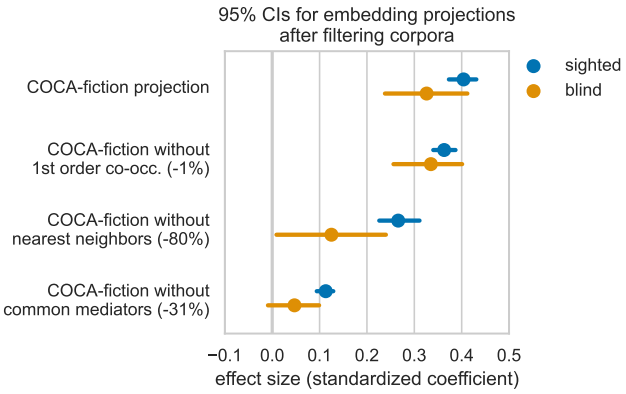

Figure 2. Estimated effects of word embedding projections in predicting blind and sighted participants' color-adjective association ratings. Percentage of training corpus removed by each manipulation is listed in parentheses.

Experiment 3

To better understand what kinds of sentences were contributing to learning human-like color-adjective associations, we modified an embedding model to record color-adjective embedding projections at every single training step. This allows us to rank training examples in order of the impact they have on specific color-adjective projections (e.g. "blue" on the axis "hot"-"cold").

Method

To measure the impact each individual sentence in the training corpus had on the embedding projections, we modified the word2vec (Mikolov, Chen, Corrado, & Dean, 2013) implementation included in the gensim Python package (Řehůřek & Sojka, 2010). The modified word2vec implementation computes and logs the embedding projections of interest after every training cycle (i.e. reading a training sentence, computing and back-propagating the error, and computing the updated embeddings). We then used the final embedding projections (after training is completed) as a reference and calculated how much each training sentence reduced the relative distance (between the previous projection and the final projection).

Results

The sentences that most informed the final embeddings projections were (1) likely to contain either a dimension word (e.g., cold) or a color word, and (2) were likely to contain a color-adjective mediator produced in Experiment 2. For example, a highly informative sentence for moving "blue" toward "cold" is"The cold seaside

air here has both a fishy and a piney sniff to it". We can count up the occurrences of color and dimension words in the top 1000 most informative sentences for the "blue" and "hot"-"cold" pairing. We find 447 occurrences of "cold", 326 occurrences of "hot", and 303 occurrences of "blue". Every sentence in the top 1000 contained at least one of these words, and only a few contained more than one. This suggests that the associations that underpin the projections are learned from specific second-order co-occurrences.

The most informative of these second-order co-occurrences are disproportionately mediated by words that participants in Experiment 2 named as salient labels for specific color associations (e.g. the association between "yellow" and "ripe" is mediated by salient label "banana"). The top 1% of informative training sentences contains 2%–6% of the participant-provided mediator words in the training corpus, when aggregated by color.

Discussion

Our results are strongly consistent with the model learning the color-adjective associations that inform the projections from second-order co-occurrence relationships. The higher prevalence of participant-proved mediator words for each color-adjective pair in the most informative training sentences demonstrates that participants were able to articulate with some degree of success which indirect (second-order) co-occurrence relationships are informative for the relationship between each given color and adjective (e.g. "white" and "cold", mediated by the word "snow").

General discussion

In a language community where word meanings are always changing and where speakers cannot observe many words' referents directly, how does a language learner align their understanding of word meanings to those of other speakers and the community at large?

One example of word meanings that have to be aligned without observing direct referents is blind people's knowledge of color words. Blind people cannot directly perceive colors in their visual contexts, yet we found that their understanding of color associations is broadly aligned with that of sighted people, and that the color associations of both groups of participants could be predicted from word embedding projections. That these color associations can be learned from a corpus of written text by a model that learns from distributional information demonstrates how media, both spoken and written, could serve to align lexical semantics across a large language community in which most members never interact directly with each other. Communicating word meaning implicitly through co-occurrence also allows a language community to incrementally develop the meanings of abstract words–for which speakers cannot make use of referents–by scaffolding them on top of more concrete words.

Here, we used an adapted word embedding model to demonstrate exactly how co-occurrence information can be used by an associative learner to learn aspects of word meaning. We show that the core signal lies in second-order linkages mediated by a third word, e.g., the link between "ripe" and "red" being mediated by "tomato". Large-scale semantic alignment and the mechanisms underpinning it are an under-explored topic in language evolution, but we believe that any comprehensive theory of language change needs to account for how language communities can maintain mutual intelligibility in the face of changing word meanings and varied access to direct perceptual information.

Acknowledgements

This research was supported by NSF BCS grant 2020969, awarded to Gary Lupyan. We would like to thank Armin Saysani and Michael and Paul Corballis for making available the raw data used in Experiment 1.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *arXiv [Preprint]*. (https://arxiv.org/abs/1607.04606)

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. *arXiv [Preprint]*. (https://arxiv.org/abs/2012.10754)

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2018). Semantic projection: Recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv [Preprint]*. (https://arxiv.org/abs/1802.01241)

Kim, J. S., Aheimer, B., Manrara, V. M., & Bedny, M. (2021). Shared understanding of color among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *118*(33).

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv [Preprint]*. (https://arxiv.org/abs/1301.3781)

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop* (pp. 45–50). (http://is.muni.cz/publication/884893/en)

Saysani, A., Corballis, M. C., & Corballis, P. M. (2021). Seeing colour through language: Colour knowledge in the blind and sighted. *Visual Cognition*, 1–9.